"Прокрустово ложе" или "испанский сапог” – мифы и реальность СУБД в Облаках (на примере ClickHouse)

Александр Зайцев, Altinity Inc





О себе

24 года в IT

CTO и основатель Altinity

Черный пояс в айкидо и ClickHouse

Не брат Петра Зайцева :)





Enterprise ClickHouse support and services

Altinity.Cloud DBaaS

https://altinity.com

Вы строите приложения, а мы готовим ClickHouse



О докладе

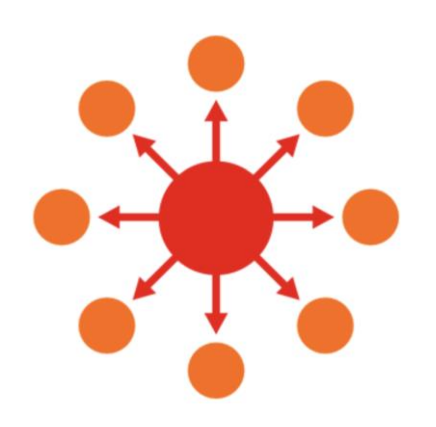
Как скрестить ужа с ежом

или

как использовать обычную СУБД в облаках



Облачные обещания и надежды









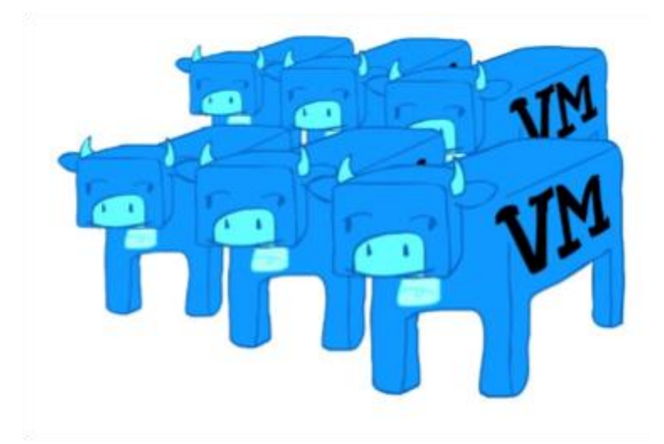


Облака – это стада (cattle)!

Стадо железных серверов

Стадо виртуалок

Стадо подов (если это Кубернетес)





СУБД – это котики (pets)



… котики с чемоданом данных



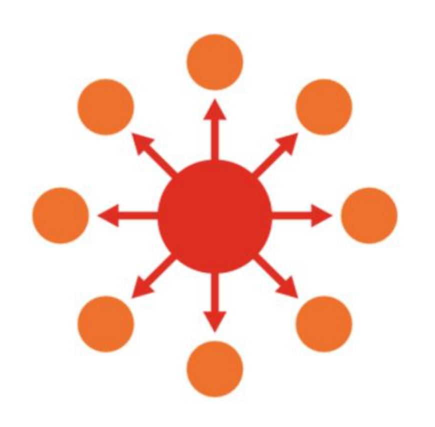


Как выполнить облачные обещания?

?

24

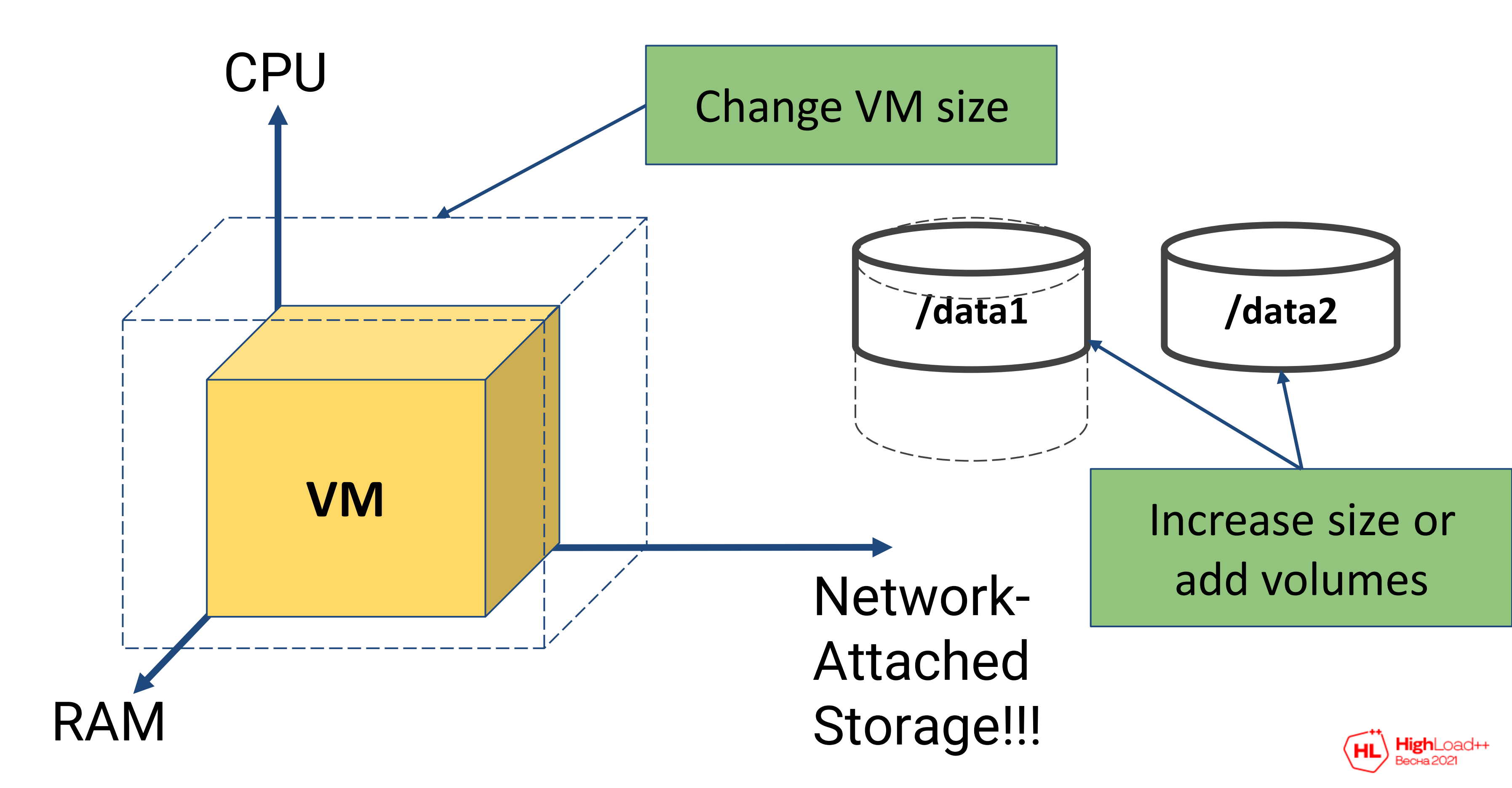




Масштабирование



Вертикальное



Требования вертикального масштабирования

К Облаку

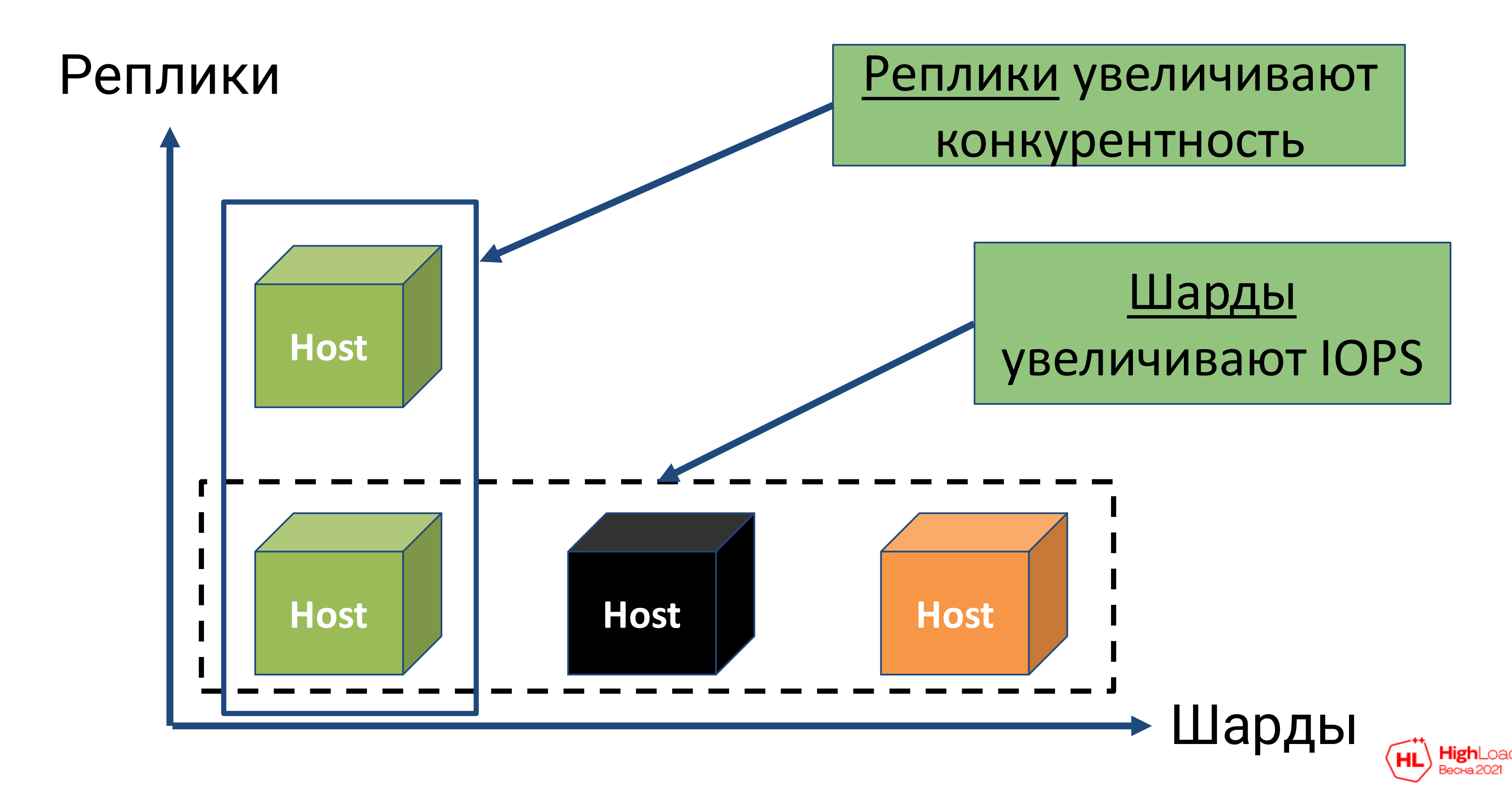
- Масштабирование томов
- Монтирование нескольких томов
- Быстрый сетевой диск

К DBMS

- Эффективно использовать RAM/CPU
- Работать с несколькими томами
- Сносно работать с «медленным» сетевым диском



Горизонтальное



Требования горизонтального масштабирования

К Облаку

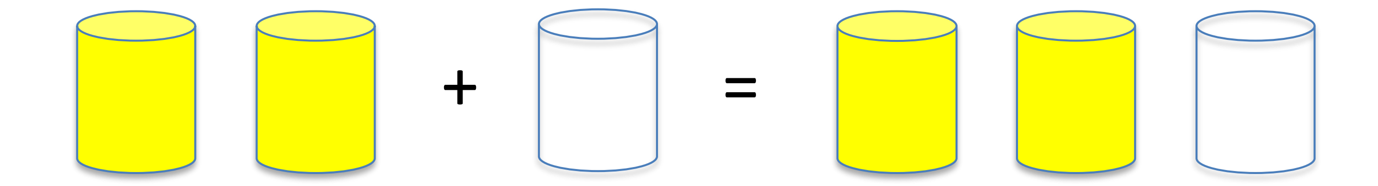
- Просто работать, идеально для «стада»

К DBMS

- Поддерживать шардирование/репликацию
- Контроль состояния репликации
- Решардинг



Репликация





Репликация

+ =



Репликация

1. Добавляем реплику, но не «включаем»
2. Ждем, пока догонится
3. Включаем в запросы



Как это сделать в ClickHouse

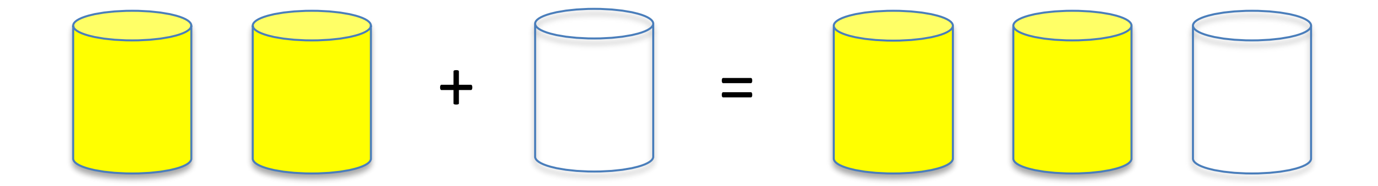
Запросы

Engine = Distributed
max_replica_delay_for_distributed_queries=10

Engine = ReplicatedMergeTree	Engine = ReplicatedMergeTree



Шардирование





Решардинг

+ =



Шардирование и решардинг

1. Добавляем шард
2. Создаем схему
3. Включаем в запросы
4. Переносим данные в фоне

Вопрос: как гарантировать целостность результатов во время переноса?



Как это сделать в ClickHouse?

- Ждать пока само выровняется
- Передвигать вручную
- Поддержка к концу 2021



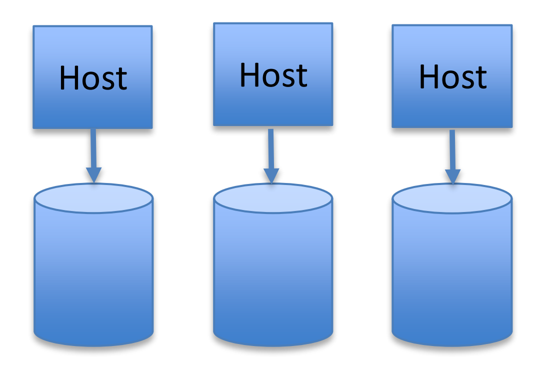
Где хранить данные в облаке?

- Локальные диски:
  - не масштабируется
  - не надежны
- Сетевые хранилища:
  - Object Storage (S3) -- хорошая скорость, но очень мало IOPS
  - Стандартные NAS (EBS gp2, gp3) – средняя скорость и цена
  - NAS с гарантированными IOPS (EBS io2 etc.) – быстро, очень дорого

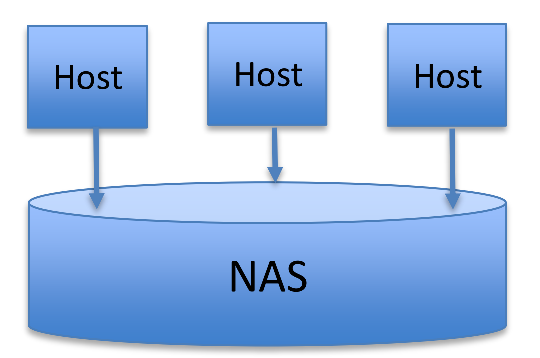


Разделение Storage и Compute

Классический bare metal



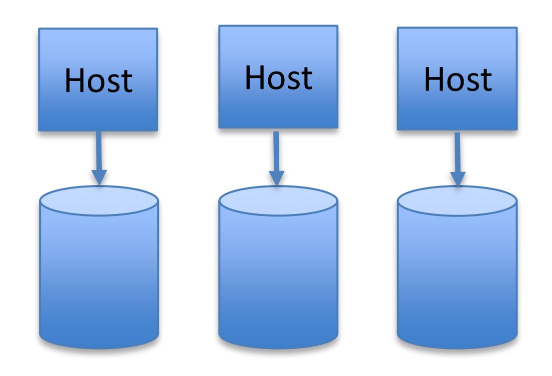
Cloud-native DBMS



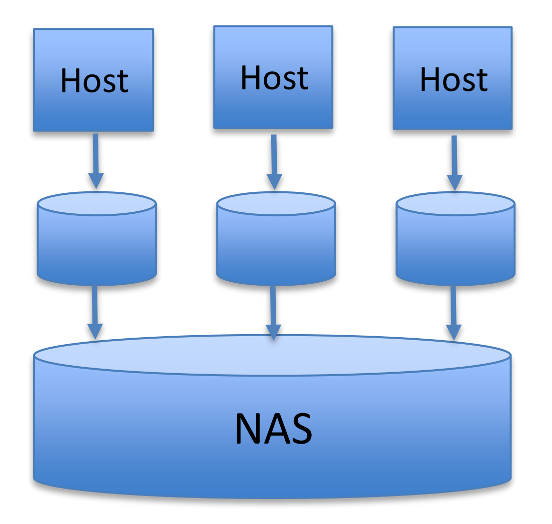


Разделение Storage и Compute

Классический bare metal



Cloud-native DBMS “на самом деле”





ClickHouse – DiskS3

EBS vs S3 MergeTree 'tripdata'

- Параллельное и multi-part чтение и запись
- Кеш индекса и «засечек»
- Zero-copy репликация в beta
- INSERT-ы медленные

https://altinity.com/blog/clickhouse-and-s3-compatible-object-storage



Архитектура хранения в ClickHouse

MergeTree tables select polices

Dimension Table A
Dimension Table B
Fact Table B

Policies define flow of data across volumes

Policy (default)
Policy

Volumes group disks into sets for writes

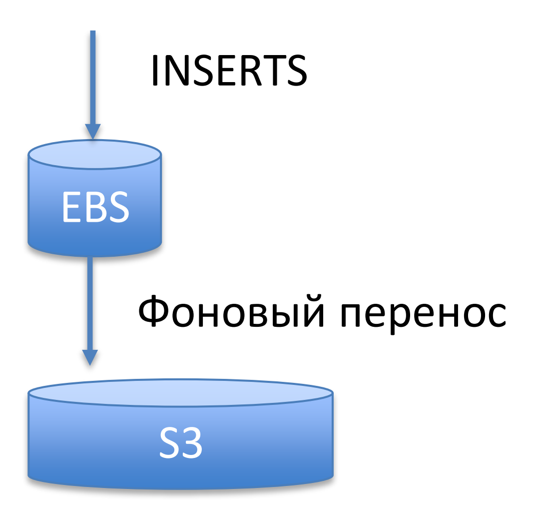
Volume
Volume
Volume

Disk entries define mount points

Disk (default)
Disk
Disk
Disk

O/S Level Devices

Tiered конфигурация:



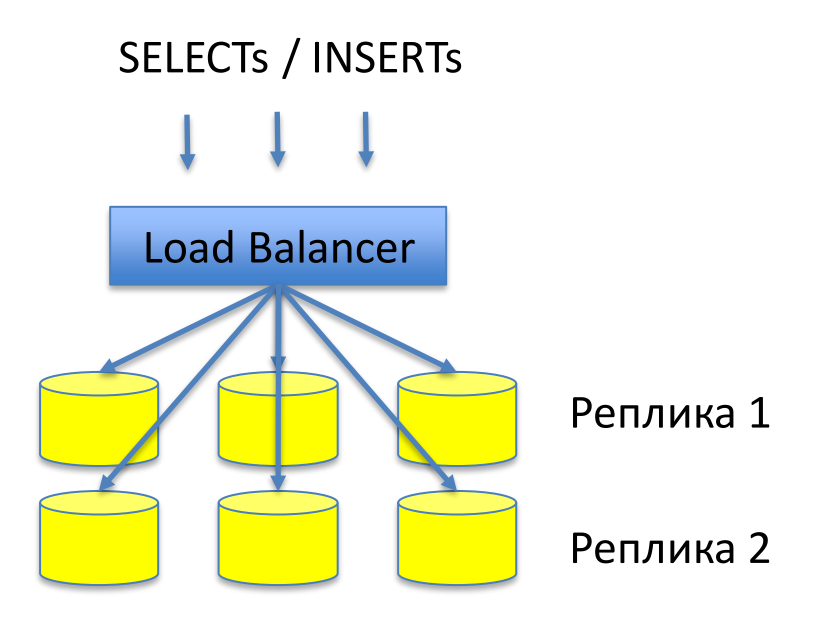




Отказоустойчивость и надежность



DBMS это не веб-сервер

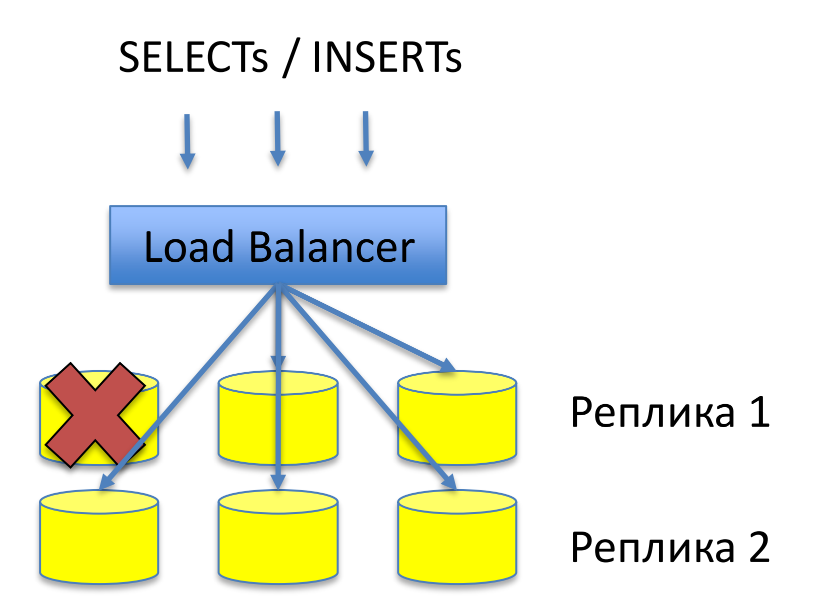


Плановый даунтайм – как сделать аккуратно?

Внеплановый даунтайм – как определить, что реплика «все»?



Плановый даунтайм



1. Убираем из балансера
2. Убираем из внутренней конфигурации
3. Ждем, пока закончатся запросы
4. Ждем, пока добежит репликация
5. Выключаем



Внеплановый даунтайм

SELECTs / INSERTs

Load Balancer

Реплика 1

Реплика 2

1. Детектим
2. Убираем из балансера
3. Разбираемся, в чем дело

Часто выздоравливает само

См: Дмитрий Столяров. «Базы данных и Kubernetes»
HL++ 2018: https://www.youtube.com/watch?v=7CR5eH6a8Fo



Облачная география

Регион US

Zone A

Zone B

Zone C

Регион EU

Zone A

Zone B

Zone C



Облачная география

Регион US

Zone A

Zone B

Zone C

Регион EU

Zone A

Zone B

Zone C



Облачная география

Регион US

Zone A

Zone B

Zone C

Регион EU

Zone A

Zone B

Zone C

Latency!



Основная ценность – данные

- Репликация в разные AZ
- Snapshots
- Backup
- Прямые руки!



Репликация между регионами

Inserts

Inserts

ZK1

Replication ->

cluster1

cluster1

<- Replication

cluster2

cluster2

ZK2

Регион US

Регион EU





Удобство



Облако – это набор инструментов



а делать все придется самим



Или привлекать «рулевого»



Облачная "матрешка"

Cloud Provider (AWS)

Private VPC

VMs

Managed k8s
(EKS)

LBs

Data



Облачный провайдер:

- Поставляет «стадо»
- Держит «хлевы» EKS

Kubernetes:

- Абстрагирует взаимодействие с провайдером
- Встроенные средства HA/FO
- User-friendly объекты



Operator = deployment + monitoring + operation

ClickHouse Operator

ClickHouseInstallation YAML file

Monitoring
Healthchecks

Kubernetes API

Load Balancer Service
Replica Service
Replica Service
Replica Service
Replica Service
Stateful Set
Pod
Persistent Volume Claim
Persistent Volume
Per-replica Config Map
Shard 1 Replica 1
Shard 1 Replica 2
Shard 2 Replica 1
Shard 2 Replica 2
User Config Map
Common Config Map

ClickHouse cluster resources

https://github.com/Altinity/clickhouse-operator



Оператор «умеет» DB в Kubernetes

Load Balancer Service

Replica Service

Replica Service

Replica Service

Replica Service

Stateful Set

Pod

Persistent Volume Claim

Persistent Volume

Per-replica Config Map

Shard 1 Replica 1

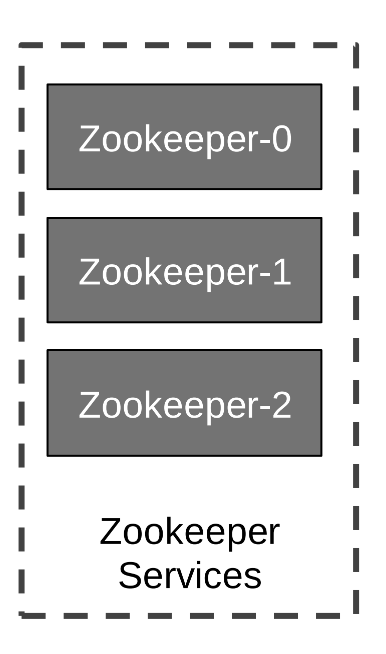
Shard 1 Replica 2

Shard 2 Replica 1

Shard 2 Replica 2

User Config Map

Common Config Map



https://github.com/Altinity/clickhouse-operator



Облачные сервисы тоже нужны

VPC

ClickHouse Client Access

AWS NLB :8443 :9443

Dashboard Browser Access

AWS NLB :443

VPC Subnet

ClickHouse

Zookeeper

Grafana

VPC Subnet

ClickHouse

Zookeeper

Prometheus

VPC Subnet

ClickHouse

Zookeeper

ClickHouse Operator

AWS EKS

EKS Control Plane

Autoscaling Group

Autoscaling Group

Autoscaling Group

S3 ClickHouse Backups

Certificate Renewals

ZeroSSL

ZeroSSL ACME Certificate Service

DNS Lookups

AWS Route 53 DNS



Взаимодействие k8s и облака

- Cluster AutoScaler
- Storage controller
- CoreDNS/ External DNS
- Аннотации сервисов
- Кастомные операторы

```
kind: Service
apiVersion: v1
metadata:
  …
  annotations:
    edge-proxy.altinity.com/port-mapping: '8443:tls-to-tls-
insecure:8443,9440:tls-to-tls-insecure:9440'
    edge-proxy.altinity.com/tls-server-name: github.demo.altinity.cloud
```

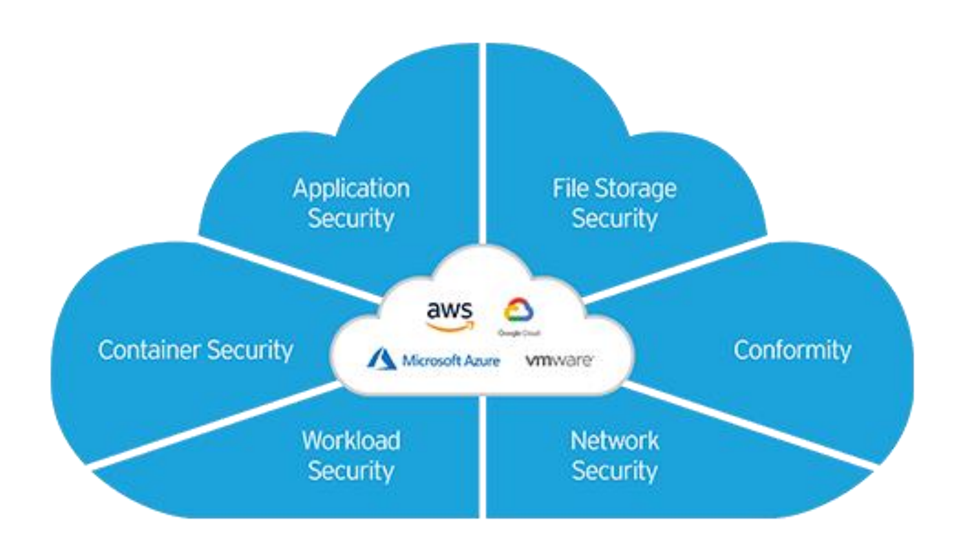


Безопасность



Облака открыты всем ветрам

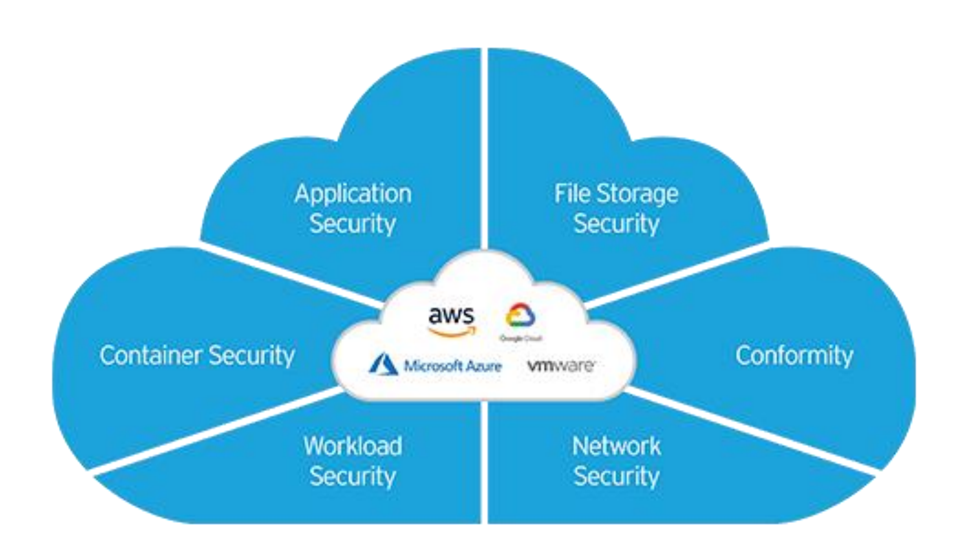




http://entradasoft.com/blogs/cloud-security-risks-and-threats-in-2020



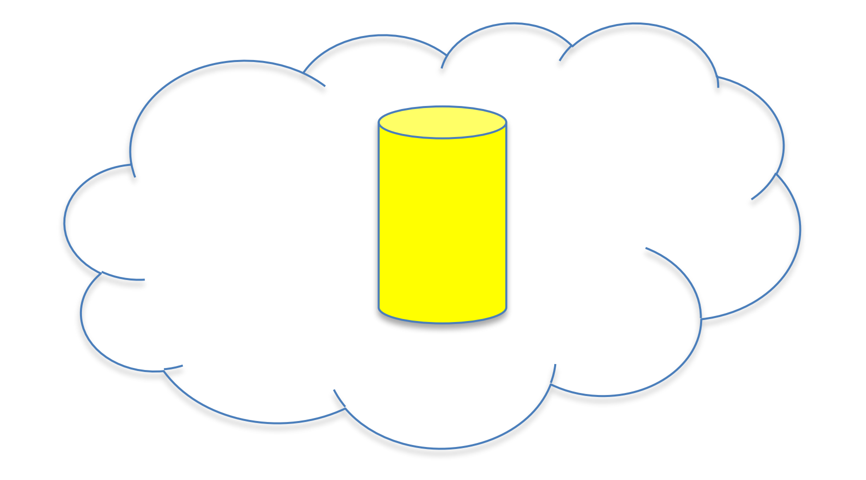
Облака открыты всем ветрам



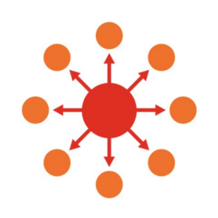
- VPC!
- Не открывать лишние порты
- IP-маски на доступ
- TLS шифрование всего трафика
- Шифрование данных
- Тестирование на проникновение
- Логгирование и Мониторинг



Резюме



- DBMS в облаках не работают «из коробки»
- Облачные обещания справедливы только для cloud-native сервисов
- DBMS приходится заворачивать в кучу доп. Софта
- Kubernetes делает жизнь в облаках проще









Спасибо!

Контакты:

alz@altinity.com

https://altinity.com

https://altinity.cloud

